DOCUMENT RESUME

ED 401 336 TM 026 151

AUTHOR	Collins, Mary A.; Chandler, Kathryn
TITLE	A Guide to Using Data from the National Household Education Survey (NHES). User's Guide.
INSTITUTION	Westat, Inc., Rockville, MD.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-96-891
PUB DATE	Sep 96
NOTE	55p.
AVAILABLE FROM	For single copies call the National Education Data Resources Center (703) 845-3151.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Adult Education; *Data Analysis; Data Collection; Discipline; Early Childhood Education; *Educational Attainment; Educational Research; *Elementary Secondary Education; *Family (Sociological Unit); National Surveys; Policy Formation; Research Methodology; *Research Utilization; School Readiness; School Safety; Telephone Surveys; *User Needs (Information)
IDENTIFIERS	Data Files; *National Household Education Survey; Random Digit Dialing; Weighting (Statistical)

ABSTRACT

The purpose of this data guide is to provide users of the National Household Education Survey (NHES) data with suggested techniques for working with the data files. Special attention is paid to topics that will help users avoid the most commonly made mistakes in working with NHES data. The NHES is a data collection system developed by the National Center for Education Statistics to provide descriptive data on the educational activities of the U.S. population. It offers policymakers, researchers, and educators a variety of statistics on the condition of education in the United States. The NHES is a telephone survey of the noninstitutionalized population of the United States for which households are selected through random digit dialing methods. The methodology for any single fielding of the NHES is linked to the research issues under study, the level of data required to address the issues, and how precise the estimates generated from the survey data need to be in order to meet study objectives. Topics addressed by NHES:91 through screening about 60,000 households were early childhood education and adult education. NHES:93 screened about 64,000 households about school readiness and school safety and discipline. The two survey components of NHES:95, early childhood program participation and adult education, paralleled NHES:91 with over 45,000 households. Three appendixes present information about commonly asked questions, data examples, and weighting and sample variance estimation. (Contains 13 appendix tables and 10 references.) (SLD)

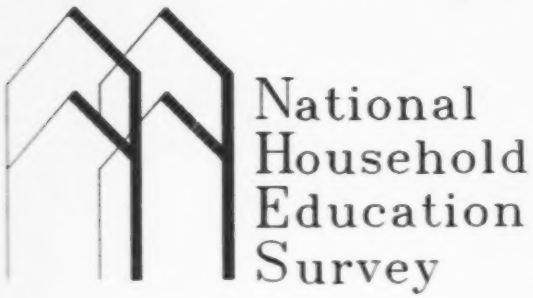
ED 401 336

NATIONAL CENTER FOR EDUCATION STATISTICS

User's Guide **September 1996**

National Household Education Survey

A Guide to Using Data from the National Household Education Survey (NHES)



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM026151

U.S. Department of Education
Office of Educational Research and Improvement **NCES 96-891**

BEST COPY AVAILABLE

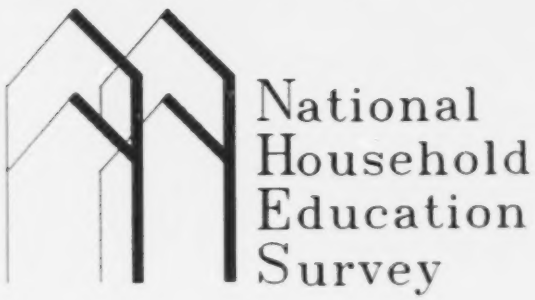
NATIONAL CENTER FOR EDUCATION STATISTICS

User's Guide

September 1996

National Household Education Survey

A Guide to Using Data from the National Household Education Survey (NHES)



Mary A. Collins, Project Director
Westat, Inc.

Kathryn Chandler, NHES Project Officer
National Center for Education Statistics

U.S. Department of Education
Office of Educational Research and Improvement

NCES 96-891

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue NW
Washington, DC 20208–5574

September 1996

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *A Guide to Using Data from the National Household Education Survey (NHES)*, NCES 96-891, by Mary A. Collins and Kathryn Chandler. Washington, DC: 1996.

Contact:
Kathryn Chandler
(202) 219–1767

For single copies, call:
National Education Data Resource Center
(703) 845–3151

Table of Contents

Section			Page
1.	**Introduction**		1
	1.1	Purpose of this Guide	1
	1.2	Overview of the National Household Education Survey	1
		1.2.1 NHES:91	1
		1.2.2 NHES:93	2
		1.2.3 NHES:95	2
	1.3	Research Issues that Can Be Addressed with the NHES Data	2
		1.3.1 NHES:91 Early Childhood Education (ECE)	3
		1.3.2 NHES:91 Adult Education (AE)	3
		1.3.3 NHES:93 School Readiness (SR)	4
		1.3.4 NHES:93 School Safety and Discipline (SS&D)	4
		1.3.5 NHES:95 Early Childhood Program Participation (ECPP)	5
		1.3.6 NHES:95 Adult Education (AE)	6
	1.4	Must Read Publications	6
	1.5	Availability of Technical Expertise	7
2.	**Brief Descriptions of the Separate NHES Data Files**		9
	2.1	NHES:91 Early Childhood Education Preprimary Data File	9
	2.2	NHES:91 Early Childhood Education Primary Student Data File	9
	2.3	NHES:91 Adult Education (AE) Data File	9
	2.4	NHES:91 Adult Education (AE) Course Data File	9
	2.5	NHES:93 School Readiness (SR) Data File	9
	2.6	NHES:93 School Safety and Discipline (SS&D) Parent and Youth Data File	10
	2.7	NHES:95 Early Childhood Program Participation (ECPP) Data File	10
	2.8	NHES:95 Adult Education (AE) Data File	10
3.	**Comparisons with Other Data Sets**		11
	3.1	Early Childhood Education and Related Issues: NHES:91 Early Childhood Education, NHES:93 School Readiness, and NHES:95 Early Childhood Program Participation	11
	3.2	Adult Education, NHES:91 and NHES:95	11
	3.3	School Safety and Discipline	12

Table of Contents--Continued

Section		Page
4.	**Getting to Know the Data**	13
	4.1 NHES Data Collection and Processing	13
	4.2 Potential Pitfalls in File Preparation	13
5.	**Selecting Variables for Working Data Sets**	17
	5.1 Developing a Research Plan	17
	5.2 Subsetting Data Files	17
	5.3 Derived Variables	18
	5.4 Subgroup Analysis	18
	5.5 Dyad Analysis	18
6.	**NHES Design**	21
	6.1 Overview of Design	21
	6.2 Household Level Sampling	21
	6.3 Topical Component Samples	23
7.	**Working with Missing Data**	25
	7.1 Sources of Missing Data	25
	7.2 Imputation Procedures in 1991	25
	7.3 Imputation Procedures in 1993	26
	7.4 Imputation Procedures in 1995	27
8.	**Weights and Estimation Procedures**	29
	8.1 Need for Special Procedures	29
	8.2 Recommended Statistical Procedures	29
	8.2.1 Software Available to Produce Weighted Estimates	29
	8.2.2 Software Available to Produce Appropriate Standard Errors	30
	8.2.3 An Alternative Method for Producing Estimates and Standard Errors	31
References		33
Appendix A: NHES:91/93/95 Commonly Asked Questions and Answers		35
Appendix B: NHES:91/93/95 Information and Examples		41
Appendix C: NHES:91/93/95 Summary of Weighting and Sample Variance Estimation Variables		53

1. Introduction

1.1. Purpose of this Guide

The purpose of this data guide is to provide users of the National Household Education Survey (NHES) data with suggested techniques for working with the data files. Special attention is paid to topics that will help users avoid the most commonly made mistakes in working with NHES data. Answers to some commonly asked questions about the NHES are also included in appendix A. This guide is meant to serve as an introduction and overview, and not as a replacement for the separate User's Manuals and other reports currently available.

1.2. Overview of the National Household Education Survey

The NHES is a data collection system of the National Center for Education Statistics (NCES). It provides descriptive data on the educational activities of the U.S. population and offers policymakers, researchers, and educators a variety of statistics on the condition of education in the United States. Although the primary purpose of the NHES is to conduct repeated measurements of the same phenomena at different points in time, one-time surveys on topics of interest to the Department of Education may also be fielded.

The NHES is a telephone survey of the noninstitutionalized civilian population of the United States. Households are selected for the survey using random digit dialing (RDD) methods. Data are collected using computer assisted telephone interviewing (CATI) procedures. The methodology for any single fielding of the NHES is linked to the research issues under study, the level of data required to address these issues, and how precise the estimates generated from the survey data need to be in order to meet the objectives of the study. The first full-scale NHES was implemented in the spring of 1991; the second and third were conducted in the spring of 1993 and the spring of 1995, respectively. Each of these collections has included two topical components.

1.2.1. NHES:91

The topics addressed by the NHES:91 were early childhood education and adult education. About 60,000 households were screened for the NHES:91. In the Early Childhood Education (ECE) component, 13,892 parents/guardians of 3- to 8-year-olds completed interviews about their children's early educational experiences. Included in this component were participation in nonparental care/education, characteristics of programs and care arrangements, and early school experiences including delayed kindergarten entry and retention in grade. In addition to questions about care/education arrangements and school, parents were asked about activities children engaged in with parents and other family members inside and outside the home. Information on family, household, and child characteristics was also collected.

In the NHES:91 Adult Education (AE) component, 9,774 persons 16 years of age and older, identified as having participated in an adult education activity in the previous 12 months, were questioned about their activities. Information collected on up to four courses included the subject matter, duration, sponsorship, purpose, and cost. A smaller sample of nonparticipants (n = 2,794) also completed interviews about barriers to participation. Information on the household and the adult's background and current employment was also collected.

1.2.2. NHES:93

In the NHES:93, about 64,000 households were screened. Nearly 11,000 parents of children aged 3 through 7 or in second grade or below completed interviews for the School Readiness (SR) component (n = 10,888). Topics included in this component were the developmental characteristics of preschoolers, school adjustment and teacher feedback to parents for kindergartners and primary students, center-based program participation, early school experiences, home activities with family members, and health status. Extensive family and child background characteristics, including parent language and education, income, receipt of public assistance, and household composition, were collected to permit the identification of at-risk children.

In the School Safety and Discipline (SS&D) component, 12,680 parents of children in grades 3 through 12 and 6,504 youth in grades 6 through 12 were interviewed about their school experiences. Topics included the school learning environment, discipline policy, safety at school, victimization, availability and use of alcohol/drugs, and alcohol/drug education. Peer norms for behavior in school and substance use were also included in this topical component. Extensive family and household background information and data about characteristics of the school attended by the child were collected.

1.2.3. NHES:95

The two survey components of the NHES:95, Early Childhood Program Participation (ECPP) and Adult Education (AE), addressed the same topics as the NHES:91. Over 45,000 households were screened for the NHES:95. In the ECPP component, parents of 14,064 children from birth through third grade were asked about their children's participation in care or education provided by relatives, nonrelatives, Head Start programs, and center-based programs. The ECPP survey also collected information on early school experiences for school-age children, home literacy activities, health and disability status, and parent and family characteristics.

For the AE interview, adults (i.e., persons age 16 and older not enrolled in secondary school) were asked about their participation in basic skills courses, English as a second language (ESL) courses, credential (degree or diploma) programs, apprenticeships, work-related courses, personal development/interest courses, and interactive video or computer training on the job. Information on programs or courses included the subject matter, duration, cost, location and sponsorship, and employer support. Nonparticipants in selected types of adult education were asked about their interest in educational activities and barriers to participation. Extensive background, employment, and household information was collected for each adult. Altogether, 19,722 adults were interviewed.

1.3. Research Issues that Can Be Addressed with the NHES Data

Each component collects specific data based on a set of research questions that guided the development of the survey component. This section lists the research questions pertinent to each component. A list of publications is also provided where appropriate. **To obtain single copies of the publications listed in this section, call (703) 845-3151**.

As the research questions shown below illustrate, a wide range of research issues can be addressed by analysts using the NHES data. However, these lists are not exhaustive. Analysts should review the survey instrument for each component to identify additional areas of particular interest to them.

1.3.1. NHES:91 Early Childhood Education (ECE)

The following research questions guided the development of the NHES:91 ECE component:

1. What developmental and education experiences do children bring to school from their homes?

2. What is the level of participation in early childhood programs, and what are the characteristics of this participation?

3. To what extent are parents delaying their children's entry into school?

4. To what extent are parents involved in their children's early childhood program participation and schooling?

5. To what combinations of educational experiences are children exposed?

6. What are the paths that children take into primary school?

7. What are the early school experiences of children?

8. What are the rates of retention in kindergarten, first grade, and second grade?

Reports using the ECE data published by NCES include: *Home Activities of 3- to 8-year-olds* (NCES Publication No. 92-004), *Experiences in Child Care and Early Childhood Programs of First and Second Graders* (NCES Publication No. 92-005), *Profile of Preschool Children's Child Care and Early Education Program Participation* (NCES Publication No. 93-133), and *Access to Early Childhood Programs for Children at Risk* (NCES Publication No. 93-372).

1.3.2. NHES:91 Adult Education (AE)

The following research questions guided the development of the NHES:91 AE component:

1. To what extent are educationally needy adults receiving basic skills and language training?

2. To what extent are adults pursuing employment- and career-related training? Are the most needy (those with low educational attainment and fewer job skills, those who have been out of the labor force) receiving training?

3. Who is providing adult education? Who pays for it? What is the involvement of business, industry, and other groups in adult education activities?

4. What is the pattern of adult education for involved adults? What is the content, format, and intensity of adult education activities?

5. Why do adults participate? What advantages do adults perceive from their participation in adult education?

6. What factors present barriers to participation in adult education activities?

NCES reports based on the NHES:91 AE component include: *Adult Education: Main Reasons for Participating* (NCES Publication No. 93-451) and *Adult Education: Employment-Related Training* (NCES Publication No. 94-471).

1.3.3. NHES:93 School Readiness (SR)

The development of the NHES:93 SR component was guided by the following research questions:

1. What is the level of parent/child interaction for young children?
2. What sources of information and advice on child development and education do parents use?
3. What is the general health/nutritional status of children?
4. In what early childhood programs do preschool children participate?
5. What early childhood programs did first and second graders participate in prior to entering kindergarten (or prior to first grade if they did not attend kindergarten)?
6. How many children experience adjustment problems when they enter kindergarten?
7. How many children experience adjustment problems in primary grades?
8. How many children are retained in kindergarten, placed in transitional grades, or retained in primary grades?

Reports that have been prepared using the SR component include the following: *Readiness for Kindergarten: Parent and Teacher Beliefs* (NCES Publication No. 93-257), *Approaching Kindergarten: A Look at Preschoolers in the United States* (NCES Publication No. 95-280), and *Family-Child Engagement in Literacy Activities: Changes in Participation Between 1991 and 1993* (NCES Publication No. 95-689). Each of these reports addresses one or more of the research questions above.

1.3.4. NHES:93 School Safety and Discipline (SS&D)

The following research questions were used to guide the development of the NHES:93 SS&D component:

1. How do parents and students perceive the learning environment of the school?
2. How safe do parents and students believe the school environment to be?
3. How is the school's discipline policy perceived?
4. To what extent are tobacco, alcohol, and/or other drug use perceived by parents and students as a problem in school?

5. Do parents and students identify alcohol and other drug education programs at school?

6. To what extent do family norms and behavior support appropriate behavior in school?

Several reports have also been prepared using the NHES:93 SS&D component. These reports include *Student Victimization at School* (NCES Publication No. 95-204), *Student Strategies to Avoid Harm at School* (NCES Publication No. 95-203), *Gangs and Victimization at School* (NCES Publication No. 95-740), *Use of School Choice* (NCES Publication No. 95-742R), and *Parent and Student Perceptions of the Learning Environment at School* (NCES Publication No. 93-281).

1.3.5. NHES:95 Early Childhood Program Participation (ECPP)

The development of the NHES:95 ECPP component was guided by the following research questions:

1. To what extent do children receive nonparental care and early childhood education?

2. Do at-risk children have the same access to nonparental care and education programs as children as a whole or as advantaged children?

3. What experiences do children have with nonparental child care or early childhood programs prior to enrolling in kindergarten or primary school?

4. How has the nonparental care or education children receive changed from 1991 to 1995?

5. To what extent do kindergartners and primary school children receive nonparental care by relatives, by nonrelatives, and in center-based or school-based before-and-after school programs, in addition to their enrollment in school?

6. How many primary school children return home from school to sibling care or self-care?

7. To what extent do children experience continuity or discontinuity in child care arrangements and early childhood programs?

8. To what extent do children in different types of arrangements have different care or education experiences?

9. How are parent expenditures associated with receipt and quality of nonparental care or education?

10. How are children's school experiences associated with receipt of nonparental care and early childhood programs prior to starting kindergarten?

11. What is the prevalence of home schooling among children of kindergarten age and in kindergarten through third grade?

12. Which aspects of nonparental care and education are most important to parents?

To date, one report has been prepared using the ECPP component and published by NCES. This report is *Child Care and Early Education Program Participation of Infants, Toddlers, and Preschoolers* (NCES

Publication No. 95-824). Additional reports are forthcoming, including one pertaining to children's delayed entry into kindergarten and retention in kindergarten.

1.3.6. NHES:95 Adult Education (AE)

The development of the NHES:95 AE component was guided by the following research questions:

1. To what extent do adults participate in educational activities, and what changes occurred in that participation from 1991 to 1995?
2. What are the main reasons for participation in adult education, and have these main reasons changed from 1991 to 1995?
3. What are the main reasons adults do not take part in adult education?
4. Who are the providers of various types of adult education?
5. How much time do adults spend in educational activities?
6. What is the involvement of employers and labor unions in the provision of adult education?
7. To what extent do adults use their own resources to participate in educational activities?
8. In what programs or courses do adults take part?
9. To what extent do adults participate in career-or job-related activities?
10. To what extent do adults use what they learn in educational activities?

To date, one report has been published by NCES for the AE component, entitled *Forty Percent of Adults Participate in Adult Education Activities: 1994-95* (NCES Publication No. 95-823). Other reports are forthcoming, including those covering the topics of participation in basic skills education, participation English as a Second Language (ESL) education, and participation in work-related education activities.

1.4. Must Read Publications

Before a researcher attempts to use the NHES data files, it is strongly suggested that time be spent reading the NHES Data File User's Manuals. The chapters of the User's Manuals can be found on the CD-ROM in WordPerfect 5.1 (.WP5) format. The following list of documents will provide researchers with comprehensive information that will help them understand the complexities of the NHES data files.

National Household Education Survey of 1991, Preprimary and Primary Data Files User's Manual (NCES Publication No. 92-057) - User's Manual covering the ECE component administered in 1991.

National Household Education Survey, Adult and Course Data Files User's Manual (NCES Publication No. 92-019) - User's Manual covering the AE component administered in 1991.

National Household Education Survey, School Readiness Data File User's Manual (NCES Publication No. 94-193) - User's Manual covering the SR component administered in 1993.

National Household Education Survey, School Safety and Discipline Data File User's Manual (NCES Publication No. 94-218) - User's Manual covering the SS&D component administered in 1993.

National Household Education Survey, Early Childhood Program Participation Data File User's Manual (NCES Publication No. 96-825) - User's Manual covering the ECPP component administered in 1995.

National Household Education Survey, Adult Education Data File User's Manual (NCES Publication No. 96-826) -User's Manual covering the AE component administered in 1995.

Additional technical publications available from NCES provide information on specific aspects of the NHES design. These include the following:

Overview of the NHES Field Test (NCES Publication No. 92-099) - A review of the major 1989 field test conducted to assess the methodology employed in the NHES.

Telephone Undercoverage Bias of 14- to 21-year-olds and 3- to 5-year-olds (NCES Publication No. 92-101) - A methodological analysis of the effects of coverage bias and poststratification procedures on survey estimates.

Effectiveness of Oversampling Blacks and Hispanics in the NHES Field Test (NCES Publication No. 92-104) - A methodological analysis of sampling procedures employed in the NHES.

As this guide is being prepared, additional technical reports are being prepared. These forthcoming publications include the following:

- *Feasibility of Conducting Followup Surveys to the National Household Education Survey*
- *Reinterview Results for the School Readiness and School Safety and Discipline Components of the 1993 National Household Education Survey*
- *Use of Cognitive Laboratories and Recorded Interviews in the National Household Education Survey*
- *Measuring Adult Education Participation*
- *Adjusting for Coverage Bias Using Telephone Service Interruption Data in the 1993 National Household Education Survey*

1.5. Availability of Technical Expertise

Staff at NCES and Westat have been closely associated with important components of the NHES design, instrumentation, sampling, data collection, data processing, and analyses that have occurred over the span of this project. These individuals, their area of expertise, and their phone numbers are included below.

Overall Knowledge of the NHES

Kathryn Chandler (NCES)	(202) 219-1767
Dan Kasprzyk (NCES)	(202) 219-1588
Mary Collins (Westat)	(301) 251-4273

Statistical Procedures

Mike Brick (Westat)	(301) 294-2004

NHES Components

NHES:91 Early Childhood Education

Kathryn Chandler (NCES)	(202) 219-1767
Mary Collins (Westat)	(301) 251-4273

NHES:91 Adult Education

Roslyn Korb (NCES)	(202) 219-1587
Peter Stowe	(202) 219-2099
Carin Celebuski (Westat)	(301) 294-3986
Kwang Kim (Westat)	(301) 517-4078

NHES:93 School Readiness

Kathryn Chandler (NCES)	(202) 219-1767
Mary Collins (Westat)	(301) 251-4273

NHES:93 School Safety and Discipline

Kathryn Chandler (NCES)	(202) 219-1767
Mary Jo Nolin (Westat)	(301) 294-2031

NHES:95 Early Childhood Program Participation

Kathryn Chandler (NCES)	(202) 219-1767
Mary Collins (Westat)	(301) 251-4273
Laura Loomis-(Westat)	(301) 517-4049

NHES:95 Adult Education

Roslyn Korb (NCES)	(202) 219-1587
Peter Stowe (NCES)	(202) 219-2099
Kwang Kim (Westat)	(301) 517-4078

2. *Brief Descriptions of the Separate NHES Data Files*

The purpose of this section is to provide a brief description of the separate NHES data files. Analysts are reminded of the importance of carefully reviewing the NHES Data File User's Manuals prior to conducting analyses of the NHES data. The User's Manuals provide extensive information on the sampling, methodology, questionnaires, data collection procedures, use of weights, and data file layouts.

2.1. NHES:91 Early Childhood Education Preprimary Data File

The NHES:91 Preprimary file contains data from 7,655 completed interviews with parents of children age 3 through kindergarten. Included are variables from the Screener and Preprimary interviews, derived variables, weights, and replicate weights. Users are able to merge this file with the NHES:91 Primary file for the purposes of conducting analyses involving all children (see the Data File User's Manual for a discussion of this procedure).

2.2. NHES:91 Early Childhood Education Primary Student Data File

The NHES:91 Primary file contains data from 6,237 completed interviews with parents of children in first grade through age 8, plus 9-year-olds in first or second grade. Included are variables from the Screener and Primary School interviews, derived variables, weights, and replicate weights. Users are able to merge this file with the NHES:91 Preprimary file for the purposes of conducting analyses involving all children (see the Data File User's Manual for a discussion of this procedure).

2.3. NHES:91 Adult Education (AE) Data File

The NHES:91 Adult file contains responses from each completed AE participant and nonparticipant interview. Included are variables pertaining to courses taken and reported by each participant respondent. In total, there are 12,568 completed Adult interviews in this file; this includes 9,774 participant and 2,794 nonparticipant interviews.

2.4. NHES:91 Adult Education (AE) Course Data File

The NHES:91 AE Course file contains information for each course reported by AE participants in the AE interview. Note that course information is also included for each adult included in the Adult File. In the Course file, the unit of analysis is each course described by adult respondents in completed AE interviews; there are 17,612 courses in the NHES:91 Course file.

2.5. NHES:93 School Readiness (SR) Data File

This file contains data from 10,888 completed SR interviews with parents of children age 3 through 7 plus children up to age 9 in second grade or below. The file is organized so that logically related sets of variables are grouped together. The data items are listed in the file in the following order: identification (or system) variables, household membership information, questionnaire item variables, derived variables, weights, replicate weights for variance estimation, and imputation flag variables.

2.6. NHES:93 School Safety and Discipline (SS&D) Parent and Youth Data File

In total, there were 19,184 completed NHES:93 SS&D interviews. Of this total, 12,680 interviews were completed by parents of third through twelfth graders and 6,504 were completed by youth in the sixth through twelfth grades. There is a separate case in the file for each interview completed; therefore there are 12,680 parent cases and 6,504 youth cases. It is also important to note that for each of the 6,504 youth cases, there is a corresponding parent case. The file is organized so that logically related sets of variables are grouped together. The data items are listed in the file in the following order: identification (or system) variables, household membership information, questionnaire item variables, derived variables, weights, replicate weights for variance estimation, and imputation flag variables.

2.7. NHES:95 Early Childhood Program Participation (ECPP) Data File

This file contains data from 14,064 completed ECPP interviews with parents of children age 10 or younger and in the third grade or below. The file is organized so that logically related sets of variables are grouped together. The data items are listed in the file in the following order: identification (or system) variables, household membership information, questionnaire item variables, derived variables, weights, replicate weights for variance estimation, imputation flag variables, and other flag variables.

2.8. NHES:95 Adult Education (AE) Data File

This file contains data from 19,722 completed AE interviews with adults aged 16 or older not enrolled in elementary or secondary school and not on active duty in the U.S. Armed Forces. Of the 19,722 adults who completed interviews, 11,713 were identified as participants in adult education activities and 8,009 as nonparticipants. Similar to the other files, logically related sets of variables in the AE data file are grouped together. The data items are listed in the file in the following order: identification (or system) variables, household membership information, questionnaire item variables, derived variables, weights, replicate weights for variance estimation, imputation flag variables, and other flag variables.

3. *Comparisons with Other Data Sets*

In addition to conducting analyses with the NHES data sets, analysts may wish to compare NHES estimates to those obtained in other large-scale data collections. When making such comparisons, researchers should carefully review methodological and wording differences that may affect responses. This section briefly describes some data sets that address issues related to the NHES components.

3.1. Early Childhood Education and Related Issues: NHES:91 Early Childhood Education, NHES:93 School Readiness, and NHES:95 Early Childhood Program Participation

The early childhood components of the NHES have addressed participation in child care and early childhood programs, home activities, health status, and a number of related issues. Data users may wish to compare some estimates, for example, center-based program participation or frequency of home activities, across NHES early childhood data sets. In addition, several extant data sources include similar items.

- The *Current Population Survey October Education Supplement* (Bureau of the Census) has collected information on enrollment in nursery school and school for many years. Since many parents report center-based early childhood programs of many types (including day care centers) in response to these items, estimates of participation in such programs can be compared. In addition, estimates of retention in early grades can be analyzed with both the Current Population Survey (CPS) and the NHES data.

- The 1990 *Current Population Survey October Education Supplement* replicated several NHES items on home activities in which parents engage with their children.

- The *National Health Interview Survey Child Health Supplement of 1988* (National Center for Health Statistics) collected information on participation in child care and early childhood education programs and extensive information on the health status of children.

- The *Survey of Income and Program Participation* (Bureau of the Census) is a recurring survey that periodically includes a supplement that collects information on the child care and early childhood program participation of children of mothers who are employed or enrolled in school or job training. Child care supplements were administered in 1991 and 1993.

3.2. Adult Education, NHES:91 and NHES:95

The NHES Adult Education component fills an important information need, since there are no other current national data available on the participation of adults in the broad range of adult education activities. As a result, researchers will not find current data available for comparison. However, those analysts interested in participation over time may wish to examine the available CPS data on this topic.

- From 1969 through 1984, the *Current Population Survey* (Bureau of the Census) included a triennial supplement on participation in adult education. Supplements were administered during the month of May in 1969, 1972, 1975, 1978, 1981, and 1984. These data can

be used to examine the participation of adults over time and in comparison with current levels of participation.

- In the 1992 *Current Population Survey October Education Supplement*, sets of items measuring adult education participation as done in previous CPS administrations and in the NHES:91 were included. The purpose of this was to assess the extent to which differences in measurement may be responsible for observed differences in participation rates.

3.3. School Safety and Discipline

- *Monitoring the Future* (National Institute on Drug Abuse) gathers information annually on the prevalence and incidence of the illicit drug use of 12th graders. In addition, it contains questions designed to describe and explain changes in many important values, behaviors, and lifestyle orientations of American youth. The survey was first conducted in 1975, and the sample was expanded in 1991 to include 8th and 10th grade students.

- The purpose of the 1989 *National Crime Victimization Survey, School Crime Supplement* (U.S. Department of Justice, Bureau of Justice Statistics) was to provide detailed information on personal crimes of violence and theft that were committed inside a school building or on school property. The basic National Crime Victimization Survey is an annual ongoing effort. The School Crime Supplement was repeated in 1995.

- The *National Education Longitudinal Study of 1988* (National Center for Education Statistics) represents a major longitudinal effort designed to provide trend data about critical transitions experienced by 8th grade students as they leave middle/junior high school and progress through high school into college or their careers. Data from this study can be used to examine educational issues such as school environment issues, school discipline issues, victimization at school, and drug and alcohol education.

4. *Getting to Know the Data*

This section of the Guide provides information to help to ensure successful use of the NHES data. Included are general discussions of the data collection and data processing procedures used in preparing the data, as well as some potential pitfalls in file preparation.

4.1. NHES Data Collection and Processing

While researchers conducting secondary analysis were not involved directly in the collection and processing of data, an understanding of these processes is important to thoughtful and appropriate use of the data. The method of data collection, the length of the interview, the process for cleaning the survey data during and after data collection, quality control activities, and the development of weights are all important considerations. For example:

- Some respondents did not complete the entire interview, but completed enough so that the interview was included as a partial complete.

- In the NHES components, as in every survey, there was some item nonresponse.

  - Some respondents did not know or could not recall the answers to specific items.
  - Some refused to answer certain questions.
  - Respondent or interviewer error led to erroneous paths that were corrected in the data preparation process based on interviewer notes.
  - Some respondents did not finish the entire interview.

The procedures used to deal with these circumstances are described in the Data File User's Manual for each component. Users will find tables showing item response rates for several questions; the item response rates presented in the tables were selected to illustrate the rates for key items, the rates for items appearing early and later in the questionnaire, and the range of item response rates. Most of the item response rates are very high in the NHES, as in most telephone surveys.

In addition to item nonresponse, most surveys contain a few data anomalies. These may be real or apparent inconsistencies in the data. These are also discussed in the Data File User's Manuals.

Analysts can perform their own quality checks on the NHES data relatively easily. A common means of doing so is comparing survey estimates to known population totals or to population estimates from well-established sources such as the Current Population Survey. Another data quality procedure, appropriate to not fully imputed data files such as the NHES:91 components, is to compare item respondents to item nonrespondents. An appropriate question the analyst would ask him or herself is "Are certain types of respondents more or less likely to refuse a question or to respond that they do not know the answer?"

4.2. Potential Pitfalls in File Preparation

Fortune and McBee (1984) have grouped pitfalls in data file preparation into seven categories. These seven categories are:

1) Sample skews;
2) Merger mortality;
3) Nonresponse noise;
4) Variant variables;
5) Aggregate anomalies;
6) Time tangles; and
7) Mechanical misuses.

In the following text, the applicable pitfalls of file preparation are defined (quoted from Fortune and McBee) and then related to the NHES.

> *Sample skews* "occur when data bases are merged, when data with missing values are used to construct variables, when nonresponse and item nonresponse are not reported, when data weights are incorrect, and when oversampling is not reported."

Analysts should review relevant sections of the Data File User's Manuals to be sure that they understand the design of the NHES, including the oversampling of some telephone clusters, the use of sample weights, and the appropriate use of replicate weights. In addition, analysts should review sections of the User's Manuals, this document (section 6), and other relevant publications in order to become familiar with the issues surrounding telephone coverage of the population.

In the construction of composite variables or indexes, the analyst should be aware of potential biases resulting from missing values. First, missing values, which are coded as negative numbers in NHES data files, should be set to missing prior to the construction of new variables. Second, users should assess the extent to which patterns of nonresponse may lead to a particular group of respondents being underrepresented in a composite index. While item nonresponse is very low in the NHES, such an examination will help to uncover any particular difficulties with missing data. This is particularly the case for the NHES:91, when full imputation of missing values was not conducted; imputation is discussed in section 7.

> *Merger mortality* "occurs when a large or disproportionate segment of a population has to be dropped from the data file on which the study is to be conducted."

Merger mortality problems typically occur in studies using multi-stage data collection designs or requiring the matching of data from different sources (e.g., transcripts and questionnaires or school and student questionnaires). The NHES data are not subject to this problem. First of all, the NHES does not collect multi-stage or longitudinal data. Also, there is only one NHES data set containing data that were collected on subjects from different sources: the NHES:93 School Safety and Discipline Component. In this component, data pertaining to sampled youth were collected from the youth themselves as well as from their parents. For these data, there should be no problems matching data from parent and youth questionnaires, because there is a completed parent interview for every completed youth interview. While the converse is not true (i.e., there is not a youth interview for every parent interview) due to subsampling of youth and youth nonresponse, this is accounted for in the sample weights for the youth.

> *Nonresponse noise* "occurs when nonresponse is systematic or disproportionate, when the reason for nonresponse cannot be determined, and when its effect cannot be estimated."

Both unit (interview) and item (question) nonresponse can introduce nonresponse noise. Unit nonresponse effects are more difficult to assess, particularly in a survey using a random digit dialing approach. Few items in the NHES have high item nonresponse rates (see Data File User's Manuals). However, when

specific types of respondents have higher nonresponse rates, or when underlying circumstances (e.g., a child's poor school performance) lead to item nonresponse, there is a potential for noise in the data.

> *Variant variables* "occur when there are slight differences in the definitions of a variable across two data bases or when the variable is coded in different ways in two data bases."

In comparing NHES estimates with estimates from other surveys, analysts may find questions asked in different formats or with different response options. In addition, some differences in the wording of NHES items may occur from cycle to cycle as part of the ongoing data quality activities of the project. Analysts should carefully check the comparability of items across surveys. Example 3 in appendix B provides an example of this type of check.

> *Aggregate anomalies* "occur when the unit of analysis and the level at which measurement occurred are different or when the organizational unit used to create a new data file is different from the one in the primary data base."

A potential for aggregate anomalies exists in the NHES when analysts combine responses from one or more children for the purpose of conducting family-level analysis. While the NHES:91 ECE component sampled all eligible children in a household, a maximum of two children per household were selected for each NHES:93 component and for the NHES:95 ECPP component. Analysts should keep this in mind when aggregating records to the household level.

> *Mechanical misuses* "can be defined as human errors that affect data processing and computer mechanics."

The complexity of the NHES data file structure provides ample opportunity for human error. This is ameliorated to some extent by editing procedures conducted online in the CATI system used to collect the data and in post-collection editing activities. However, the potential for such error can never be eliminated.

Mechanical misuses can also occur in the analytical stage of data processing. For example, if a person were to run frequencies for the entire School Safety and Discipline data file and report a percentage based on all records, this would be a mechanical misuse because it does not take into account that there are both parent and youth records contained in the file.

5. *Selecting Variables for Working Data Sets*

5.1. Developing a Research Plan

After a researcher 1) understands how the NHES data were collected and processed, 2) avoids the common pitfalls in data file preparation, 3) understands the limitations of the data, and 4) studies the research issues that can be addressed, he/she is ready to begin developing a research plan.

The working data file will be used by the researcher to test the research questions that are derived from previously conceived conceptual models. Before a working data set is created, the following steps are suggested:

1. Develop a research question -- What does prior research suggest is happening (e.g., how do parents and students perceive the learning environment of the school)?

2. Determine the predictor, or independent, variables (e.g., school grade level, school type, school size, student's race/ethnicity, and parents' highest education) and outcome, or dependent, variables (e.g., academic challenge, enjoyment of school, mutual respect between pupils and teachers, good discipline maintained by teachers and administrators, peer norms that support hard work for achievement, and peer norms that support good behavior) that can be used to answer the research question.

3. Determine what aspects of the research question can be answered with NHES data. If there are multiple sources of the data (e.g., parent and youth responses) available, decide if only one source or both would be the most analytically appropriate.

4. Rethink the original research question. If the variables contained on the NHES data files cannot be used to study your original research question, rethink the research question and either modify the research question or choose another data set.

5.2. Subsetting Data Files

Once the above steps have been completed, it is time to create your working data file containing only the variables and sample you are interested in examining. The following steps to subsetting are suggested:

1. According to your research question, determine which variables are needed from the NHES data file.

2. Determine whether your analysis calls for the subsetting of the data by population, e.g., adults in the labor force, children of mothers in the labor force, or youth in grades 9 through 12.

3. Use SAS or SPSS to select the appropriate cases and variables from the file. This is easily done using the NHES:91/93/95 Electronic CodeBook (ECB) which permits users to generate SAS or SPSS code that will create an extract data file that includes a subset of the cases and/or variables included in an NHES data file. See the ECB User's Guide for additional information. Also see example 1 in appendix B for an illustration of how to check that a data subset was created correctly.

5.3. Derived Variables

Derived variables were developed and included in the public use data file to aid users in analysis. The derived variables fall into three categories: questionnaire item variables, counter variables, and variables linked to other data sources (only applicable to NHES:93 and NHES:95). Questionnaire item-derived variables were created by combining two or more items from the questionnaire. Counter-derived variables were created by counting the number of persons enumerated in the household with specific characteristics. In the NHES:93 and NHES:95, linked-derived variables were created by using the respondent's ZIP Code or telephone number to extract data from other data sources, most notably the 1990 Census of Population Summary Tape File 3B (STF3B). The derived variables are on the file in alphabetical order.

In the NHES:93 School Safety and Discipline file, most derived variables come from parent responses and are included on the youth record to expedite analysis. The exceptions are FEARP, KNOWP, VICTIMP, and WITNESSP, which were derived from parent responses and are found on the parent record only, and FEARY, KNOWY, VICTIMY, and WITNESSY, which were derived from youth responses and are on the youth record only.

In the NHES:93 and NHES:95, all of the variables that begin with the prefix ZIP were taken from the 1990 Census of Population STF3B. All unique NHES:93 and NHES:95 ZIP Codes were matched to ZIP Codes on the STF3B for urbanicity, percent black or Hispanic, and percent of persons under age 18 living in poverty.

5.4. Subgroup Analysis

The NHES is designed to support examination of some specific policy-relevant subgroups. One such subgroup is minorities. Hispanics and blacks were selected at a higher than normal rate (oversampled) in order to improve national estimates. The data file can be subsetted to include only the specific group or groups of interest. In the Electronic CodeBook (ECB) for the NHES:91 through the NHES:95, special screens are provided in the data extraction menu that permit users to subset the data using a small number of variables that are commonly of interest. See the ECB User's Guide for additional information.

Data users should be aware that while poststratification techniques were found to eliminate the vast majority of bias that may be due to telephone undercoverage, it is not entirely possible to do so. Adults who have not finished high school and households with very low incomes are less likely than others to live in telephone households. Therefore, users should carefully consider possible bias when conducting analyses that focus on those with very low income. Telephone coverage bias and techniques used to correct for it are discussed in coverage reports cited in section 1.4.

5.5. Dyad Analysis

The NHES:93 SS&D component permits the analyses of data from parent and youth dyads, that is, parents and youth from the same household who completed parent and youth interviews, respectively. The variables PARNYUTH and MAINRSLT in the NHES:93 SS&D data file should be used to identify parent and youth cases that have associated youth and parent interviews. To analyze data directly comparing a parent and his/her youth's responses in the NHES:93 SS&D component, the analyst should rename the variables from one interview so as not to cause overwriting of values when the records are merged. For example, both parent and youth records contain the variable SSSTEAL. The values on the

parent records are from parent responses to the question, and the values on the youth record are from the youth responses to the question. For an analysis that includes both parent and youth, the variable name should be changed so it will be unique for parent and youth. For example, SSSTEAL could be changed to SSSTEALP on the parent record and SSSTEALY on the youth record. If the original variable name has eight characters, the P or Y will replace the last character. Because each variable will then have a unique name, the original values will remain when parent and youth records are merged.

6. NHES Design

6.1. Overview of Design

The NHES was developed to provide reliable estimates for the different topical components of each administration. The inclusion of two survey components made the overall survey more cost effective, thus allowing for larger sample sizes and more precise estimates. This strategy was key to the NHES design. By including more than one topic within the framework of a single survey, the cost of screening households to find those eligible for the study could be partitioned over the component surveys.

Another general feature of the NHES was developed in response to concerns about the potential demands placed upon those who respond to multiple survey components. With the introduction of multiple surveys within a single framework, the possibility of increasing response burden on the members of the sampled households arose. It is possible that the same household member could be selected to respond to more than one interview and/or that more than one household member could be sampled.

Even though sampling methods reduced the number of interviews per household, the length of the interview was considered to be a critical factor in obtaining high response rates and reliable estimates. Therefore, the number of items included in the NHES was limited in order to help improve response rates and reduce the demands made on survey respondents.

Because of the above requirements, complex sampling techniques, and the need for quick and accurate administration, the NHES was conducted using CATI technology. Some of the advantages of CATI for the NHES included improved project administration, online sampling and eligibility checks, scheduling of interviews according to a priority scheme to improve response rates, managing data quality by controlling skip patterns and checking responses online for range and consistency, and an online "help" function to answer interviewers' questions.

Several different interview instruments were used in each cycle of the NHES. For both the NHES:91 and NHES:93, three instruments were used; for the NHES:95 four instruments were used. These instruments included screening interviews and the extended interview topical components. Items within each of the instruments were programmed so that the appropriate items appeared on the interviewer's computer screen corresponding to the respondent's answers to previous queries.

6.2. Household Level Sampling

The sampling method used for the NHES:91 and NHES:93 is a variant of random digit dialing (RDD) procedures described in Waksberg (1978). The original Mitofsky-Waksberg method produces an equal probability sample of households with telephones and requires a smaller number of telephone calls than the sampling procedures previously used for RDD. A time-saving variant of this method, referred to as the "modified Waksberg procedure," was used for both the NHES:91 and NHES:93. The modified method is described in Brick and Waksberg (1991). Beginning in 1995, the NHES moved to a list-assisted sampling approach (Casady and Lepkowski 1993). This method reduces the number of unproductive calls to nonworking or nonresidential numbers (compared with simple random sampling of all numbers), produces a self-weighting sample, is a single stage and unclustered sample, and eliminates the sequential difficulties associated with the Mitofsky-Waksberg method. However, a disadvantage of the list-assisted method is that it incurs a coverage bias because not all telephone households are included in the sampling frame.

For the NHES:91 and NHES:93, the first step in the sampling process was to form a list of all existing telephone area codes and prefix numbers for the 50 states and the District of Columbia (a prefix number is a 3-digit telephone exchange). The lists used for the NHES samples were the Bellcore tapes for the October preceding data collection. All possible combinations of 2-digit numbers were then added to these numbers to create a list of all the possible first 8 digits of the 10 digits in telephone numbers. These 8-digit numbers were treated as Primary Sampling Units (PSUs), or telephone clusters.

A random sample of PSUs was selected. A prime telephone number was formed by adding a random two-digit number to the eight-digit cluster. The prime number was then dialed to determine if it was residential. If it was residential, the PSU was retained in the sample. If the prime number was not residential, then the PSU was rejected and no further calls within the PSU were made. Additional PSUs were selected in the same way.

A random sample of telephone numbers within each of the retained "residential" PSUs was selected by adding random two-digit combinations to the original eight numbers. Interviews were attempted at the prime number and at as many additional numbers required to obtain the desired expected sample size. The total expected sample size was $m(k+1)$, where m was the number of residential PSUs and $(k+1)$ was the number of telephone numbers sampled in each PSU.

The households were sampled within clusters in order to effect a significant cost savings. With this method of cluster sampling, the number of telephone numbers that need to be dialed is at least 50 percent less than what would be needed if all telephone numbers were dialed at random. However, the variances of the estimates were increased slightly due to the clustering of the sampled households within the PSUs.

The sampling method for the NHES:91 and NHES:93 used a fixed number of telephone numbers per PSU, rather than a fixed number of households per PSU, as used in the Mitofsky-Waksberg method. The statistical properties of this method are described in detail by Brick and Waksberg (1991). The main advantage of this method is that it does not require sequential modification to the within-PSU sample size.

The list-assisted sampling used in the NHES:95 was conducted by stratifying telephone numbers by the type of 100-bank they fall within (all the numbers in a 100-bank have the same first 8 digits of the 10-digit telephone number). An equal probability random sample of telephone numbers was selected from all telephone numbers that were in 100-banks with at least one White Page directory-listed telephone number (called the listed stratum). Telephone numbers in 100-banks with no listed telephone numbers (called the zero-listed stratum) were not sampled. The telephone numbers in the listed stratum included both listed and unlisted numbers provided there was at least one telephone number in the 100-bank that was listed.

With the list-assisted approach, a coverage bias arises because households in the zero-listed stratum have no chance of being included in the sample. Empirical findings were presented by Brick, Waksberg, Kulp, and Starer (1995) to address the question of coverage bias. These results show that the percentage of telephone numbers in the zero-listed stratum that are residential is very small (about 1.4 percent), and about 3 to 4 percent of all telephone households are in the zero-listed stratum. Furthermore, the bias resulting from excluding the zero-listed stratum is generally small.

One of the goals of the NHES is to produce reliable estimates for subdomains defined by race and ethnicity. In fact, estimates by race and ethnicity were key in developing the sample sizes for each of the administrations of the NHES. In a 64,000-household design in which every household has the same probability of being included, the number of completed interviews would not be large enough to produce

reliable estimates of many characteristics of black and Hispanic youth. Therefore, blacks and Hispanics had to be sampled at higher rates to improve the reliability of estimates for these subpopulations.

In each administration of the NHES, exchanges with higher concentrations of blacks and Hispanics have been oversampled. Mohadjer and West (1992) showed that this method was successful in reducing the variances for estimates of characteristics of blacks and Hispanics by approximately 20 to 30 percent over a range of statistics examined. The decreases in precision for estimates of the groups that were not oversampled and for estimates of totals were modest, ranging from about 5 to 15 percent.

A computer file containing census characteristics for telephone exchanges was used to stratify telephone exchanges into low- and high-minority concentration strata. Any telephone exchange not found on the file was assigned to the low-minority concentration stratum.

The specific design defined high-minority concentration areas as exchanges having at least 20 percent of black or Hispanic persons (or Asian/Pacific Islander persons for the NHES:93) living in the area. The telephone exchanges in the two strata were identified and a systematic sample was drawn in each stratum. The sampling fraction used in the high-minority concentration stratum was two times the fraction used in the low-minority concentration stratum.[1]

Oversampling by the characteristics of the telephone exchange had two effects. First, the oversampling increased the sample sizes for minorities because they were more heavily concentrated in the exchanges that were oversampled. Therefore, the sampling errors for estimates of these groups were reduced due to the increased sample size. On the other hand, not all minorities were found in the oversampled exchanges. Thus, differential sampling rates were applied to persons depending on their exchanges. Using differential rates increased the sampling errors of the estimates. These increases partially offset the benefit of the larger minority sample sizes.

6.3. Topical Component Samples

The NHES:91, NHES:93, and NHES:95 each included two survey topics. The NHES:91 AE sampled households for adults (age 16 years or older) who participated in at least one adult education course or activity during the past 12 months (including full-time, degree-seeking students) and adults who had not participated in the past 12 months. For the ECE survey, two specific populations of 3- to 8-year-old children were sampled: those who had not yet enrolled in primary school and those who were currently enrolled in primary school (or were 6 years of age or older and receiving home schooling or education in alternative programs). In addition, 9-year-olds who were enrolled in second grade or below were also included to improve estimates of retention in early grades. The age-eligibility was determined by the child's age on December 31, 1990.

In the NHES:93 SR component, data were collected about children aged 3 through 7 or in second grade or below. Those children who were 8 or 9 years old, but who were enrolled in first or second grade, were also eligible for the SR survey. Also, those children who were 7 or younger but enrolled in third grade qualified for this survey. The age-eligibility was determined by the child's age on December 31, 1992.

---

[1]Research was done for the NHES Field Test of 1989, the NHES:91, and the NHES:93 that tested the effects of different sampling plans and definitions of high minority strata on sample sizes and variances of estimates. This research led to implementing the procedures just described for oversampling telephone numbers in high minority areas.

The SS&D component of the NHES:93 was designed to gather information about the school environment, safety at school, school discipline policy, and alcohol/other drug use and education. The respondent parent or guardian was interviewed about his/her sampled child who was enrolled full time in any grade 3 through 12. The youth was usually between 8 and 20 years old, as determined by his/her age on December 31, 1992. (Youths who were age 21 and enrolled in 12th grade or below were sampled at the Screener. If, at the beginning of the extended interview, it was determined that the youth was over 20 on December 31, 1992, the interview was terminated.) A subsample of youth in 6th through 12th grades, generally age 11 and older, were also interviewed about their school experiences. There were some interviews with "emancipated youth," that is, youth who were enrolled full time in 6th through 12th grade and living independently from any parent/guardian. There are relatively few emancipated youth cases (n=77); all were 16 to 20 years old and enrolled in 10th through 12th grades. These youth were asked the same questions regarding school experiences as other youth, plus some questions pertaining to child, family, and household characteristics that normally would have been answered by the parent/guardian.

In the NHES:95 ECPP component, data were collected about children aged 10 or younger and in third grade or below. The age-eligibility was determined by the child's age on December 31, 1994. For the NHES:95 AE component, adults age 16 or older and not enrolled in elementary or secondary school and not currently serving in the U.S. Armed Forces were sampled for interviews. The probability of an adult being sampled depended upon whether he or she had a high school diploma and whether he or she had taken any classes in the previous 12 months.

7. *Working with Missing Data*

7.1. Sources of Missing Data

As in most surveys, the responses to some data items are not obtained for all interviews. There are numerous reasons for item nonresponse. Some respondents do not know the answer for the item or do not wish to respond for other reasons. Some item nonresponse arises when an interview is interrupted and not continued later, leaving items at the end of the interview blank. Item nonresponse may also be encountered because responses provided by the respondent are not internally consistent, and this inconsistency is not discovered until after the interview is completed. In these cases, the items that are not internally consistent were set to missing.

For most of the data items collected in the NHES, the item response rate was very high. In the NHES:91, missing data were imputed for those variables required for weighting or contributing to the derived variables. In the NHES:93 and NHES:95, all of the data items with missing data on the file were imputed. Thus, for the NHES:93 and NHES:95 the only missing values remaining are those that indicate legitimate skips (see example 2 in appendix B for an illustration of this). The imputations were done for two reasons. First, certain variables were used in developing the national estimates and complete responses were needed for this purpose. These included the variables used for raking and for developing sampling weights. Second, some data items were expected to be analytical variables in many of the publications from the surveys and complete item responses helped to improve these presentations.

7.2. Imputation Procedures in 1991

As stated above, the NHES:91 was the only NHES survey for which only a subset of the data items were imputed. For variables that were imputed in the NHES:91, a nearest-neighbor, hot-deck procedure was used to impute missing responses. In this approach, the entire file was placed into a specified sort order that varied depending on the data items to be imputed. The sort order was determined by attempting to group respondents into those most likely to have the same response for the data item to be imputed. For example, consider imputing the variable DADGRADE, the highest level of education completed by the child's father. The sort variables for the imputation were the child's parents' marital status (PARNMARI), the mother's education level (MOMGRADE), and whether the father was employed (DADWORK). The use of these sort variables in combination assured that adjacent cases on the file were similar on all of these characteristics.

Whenever a case with a missing value was encountered, the value of the data item from the preceding complete case was imputed for the missing item. The method is called a nearest-neighbor, hot-deck approach because the value from the closest record (the nearest-neighbor) in the current data set (hot-deck) is used to replace the missing item. Thus, in the example above, the value for a missing DADGRADE was imputed from a case with the same responses for the sort variables (PARNMARI, MOMGRADE, and DADWORK) whenever possible.

For each imputed data item, an imputation flag variable was created. If the response for this item was imputed, then the imputation flag was set equal to 1, otherwise it was set to -1 (inapplicable). The flag was created to enable users to identify imputed values. Users may use the imputation flags to delete the imputed values, use alternative imputation procedures, or account for the imputation in computation of the reliability of the estimates produced from the data set.

7.3. Imputation Procedures in 1993

In the NHES:93, a slightly different hot-deck procedure was used to impute missing responses. In this approach, the entire file was sorted into cells defined by characteristics of the respondents. The variables used in the sorting were general descriptors of the interview and also included any variables involved in the skip pattern for the items. This portion of the procedure was very similar to the sorting used in the NHES:91.

All of the observations were sorted into cells defined by the responses to the sort variables, and then divided into two classes within the cell depending on whether or not the item was missing. For an observation with a missing value, a value from a randomly selected donor (observation in the same cell but with the item completed) was imputed for the missing value. After the imputation was completed, edit programs were run to ensure the imputed responses did not violate edit rules.

Because editing was being finished at the same time as imputation was occurring, there were some logically inconsistent values, newly missing values, and imputed values that were out of range. These values were set to missing during editing. Further imputations were then necessary. A simplified manual imputation was used for these missing values because there were so few imputations recoded. The distribution of the completed data was used to draw donors for this manual process. Thus, for these newly missing values, the standard sort variables were not used to control the process.

The general imputation procedures were not used for several variables that were collected only once per household or involved complex relationships. The AGE1 through AGE9, SEX1 through SEX9, RELATN1 through RELATN9 (only on the SR file), and CRELN1 through CRELN9 (only on the SR file) household membership items were manually imputed for the very few cases that had missing values because of the need to ensure consistency in household relationships.

ZIP Code values were imputed once at the household level and then included in the SR and the SS&D files. Some 263 households were missing the ZIP Code, and another 12 households gave ZIP Codes that did not match ZIP Codes on the 1990 Census of Population STF3B used to create derived variables. These ZIP Codes, which affected 390 interviews, were imputed by replacing them with ZIP Codes that were on the STF3B file.

When the hot-deck imputation procedures were completed, the 3 SR items and 12 SS&D items with response rates of less than 95 percent were further examined. A search was conducted to find correlated variables that could be used in place of the standard sort variables for these items. If useful correlates were identified, they were used in the hot-deck imputation for these items.

For each data item for which any values were imputed, an imputation flag variable was created. If the response for the item was imputed, the imputation flag was set equal to 1 (or 2 -- see next paragraph); otherwise it was set to 0. There were no imputation flags created for AGE92, GRADE, DPAFRAID, and HNPUBL4 since there was no imputation done for these variables. The flag was created to enable users to identify imputed values.

Analysts of the SS&D data file might find a "don't know" response to be analytically useful for some items from the parent interview. A parent response of "don't know" may indicate lack of interest or involvement in the child's school experiences, whereas a youth response of "don't know" to the same question has different implications. To support this analytic objective, the imputation flag was set to the value 2 for a "don't know" response that was imputed for selected parent items. A list of these items is contained in the NHES:93 School Safety and Discipline Data File User's Manual.

7.4. Imputation Procedures in 1995

The imputation procedures used in the NHES:95 were similar to those used in the NHES:93. A hot-deck procedure was used in which the entire file was sorted into cells defined by characteristics of the respondents. The variables used in the sorting were general descriptors of the interview and also included any variables involved in the skip pattern for the items.

All of the observations were sorted into cells defined by the responses to the sort variables, and then divided into two classes within the cell depending on whether or not the item was missing. For an observation with a missing value, a value from a randomly selected donor (observation in the same cell but with the item completed) was imputed for the missing value. After the imputation was completed, edit programs were run to ensure the imputed responses did not violate edit rules.

For some items, the missing values were imputed manually rather than using the hot-deck procedure. This happened most often when the variable was collected only once for the household or involved complex relationships (e.g., variables indicating the ages and relationships of household members). Manual imputation was also used if edit failures were found after the hot-deck imputations were completed.

Some additional measures were taken to impute variables that had item response rates of less than 95 percent. To improve the imputation for these items, additional sort variables were added to the standard sort variables for the hot-deck imputation. Additional sort variables were also included in the hot-deck imputation of age-related variables (e.g., the age at which children first started attending a center-based program).

For each data item for which any values were imputed, an imputation flag variable was created. If the response for the item was not imputed, the imputation flag was set equal to 0. If the response was imputed, the flag was set to either 1, 2, or 3. An additional code of 4 was utilized for some imputation flags in the ECPP file. The value of the imputation flag indicates the specific procedure used to impute the missing value. The imputation flag was typically set to 1 if the missing value was imputed using the standard hot-deck approach. In some cases, variables had to be recoded to be consistent with the skip patterns of the questionnaire prior to being imputed using the standard hot-deck approach. For these cases the imputation flag was set to 2. For some items with complex skip patterns and only a few missing values, the item was imputed manually and the flag was set to 3. In the ECPP file, the imputation flag was set to 4 if the reported value was "don't know" prior to imputation using the standard hot-deck approach. Code 4 was utilized for only a subset of ECPP variables for which a "don't know" response might be considered analytically meaningful, specifically, items concerning parent knowledge of care provider or program characteristics.

The imputation flags were created to enable users to identify imputed values. Users can employ the imputation flag to delete the imputed values, use alternative imputation procedures, or account for the imputation in computation of the reliability of the estimates produced from the data set. For example, some users might wish to analyze the data with the missing values rather than the imputed values. If the flag corresponding to the variable is not equal to 0, the user can replace the imputed response with a missing value to accomplish this goal. This method could also be used to replace the imputed value with a value imputed by some user-defined imputation approach. Finally, if the user wishes to account for the fact that some of the data were imputed when computing sampling errors for the estimates, the missing values could be imputed using multiple imputation methods (Rubin 1987) or imputed so that the variance procedures in Rao and Shao (1992) could be used.

8. *Weights and Estimation Procedures*

8.1. Need for Special Procedures

In most standard statistical textbooks and software, the analyst assumes that the data are a simple random sample from some distribution. In the NHES and most other sample surveys, this assumption is incorrect because the sampled units are selected using techniques such as clustering, stratification, and unequal probabilities of selection. These sampling methods are used because they greatly reduce the cost of data collection and produce efficient and unbiased estimates of the population if the appropriate methods are used in the analysis stage.

In all NHES sample designs, telephone numbers were stratified and selected with unequal probabilities of selection. Persons within the households were stratified and selected with varying probabilities. Estimation techniques, reflected in the sampling weights, were used to make the estimates from the NHES consistent with population control totals and these estimation methods also served to reduce the variability of the estimates and remove some of the potential biases in the estimates, especially bias due to undercoverage.

All of these sampling and estimation techniques have consequences for the analytic methods that should be applied in making estimates from the NHES data. One of the most important features is that the sampling weights should always be used when making estimates of the population. These weights are important not only for estimates of totals (unweighted procedures are not reasonable for estimates of totals), but also for estimates of means and proportions. As noted above, the sampling and estimation procedures used in creating the sampling weights adjusted for the most important sources of biases, and unweighted methods of estimating do not include these adjustments.

The sampling and estimation procedures also have an important impact on the estimates of the reliability of the estimates from the NHES. The standard errors of the estimates (or the variance of the estimate which is just the square of the standard error) are affected by these procedures. If the standard errors are computed using standard statistical software such as SAS or SPSS, they will underestimate the actual standard errors for most estimates because the data are not a simple random sample (the data are correlated and sampled with unequal probabilities).

The role of the sample design and estimation procedures in computing estimates and the standard errors of the estimates is discussed in a more general setting by Kish (1992). All analysts are urged to take these features of the NHES into account when producing estimates, whether they are simple estimates of means or more complex estimates of correlations and regression coefficients.

8.2 Recommended Statistical Procedures

The recommended methods for producing estimates and their standard errors from the NHES are discussed below.

8.2.1 Software Available to Produce Weighted Estimates

Most standard statistical software has the capability of handling sample survey weights in calculating estimates. (The table in appendix C shows the full sample weight variable for each NHES data file.) For example, in SAS the WEIGHT statement can be used with SAS procedures to produce estimates

using the sampling weights. While the estimates of the characteristics may be appropriate, the standard errors of the estimates will not be correct because the procedures assume that the data are from a simple random sample. The WesVarPC and SUDAAN software packages (described below in 8.2.2) are able to compute appropriate standard errors along with weighted estimates.

8.2.2 Software Available to Produce Appropriate Standard Errors

The preferred method of producing unbiased estimates of the population and valid estimates of the standard errors of the estimates is to use software designed with this purpose in mind. The two major methods of proceeding are by using either replication methods or Taylor series approximations. Special software is available for both methods and the NHES data supports either type of analysis.

We recommend the replication method for the NHES. The replication method involves splitting the entire sample into a set of groups, or replicates, based on the actual sample design of the survey. The survey estimates can then be estimated for each of the replicates by creating replicate weights that mimic the actual sample design and estimation procedures used in the full sample. The variation in the estimates computed from the replicate weights can then be used to estimate the sampling errors of the estimates from the full sample. Replicate weights have been included in all the NHES data files to make this application relatively simple. The table in appendix C lists the replicate weight variables, as well as the specific jackknife replication method used to create the replicates, for each NHES data file.

WesVarPC is software developed for the PC for producing estimates and their standard errors using replication methods. The replication method is especially useful for the NHES because this is the only method that accounts for both nonresponse adjustments and the raking adjustments to the population control totals in the estimation of the standard errors. WesVarPC currently supports a wide variety of estimates (totals, means, proportions, ratios, and user-defined functions of estimates) as well as procedures for estimating linear and logistic regression coefficients. WesVarPC can read SAS (version 604), SAS Transport, and SPSS for Windows system data files[2].

The WesVarPC software is available free of charge through the Internet (http://www.westat.com) or by sending an e-mail message to wesvar@westat.com. Those interested in obtaining a copy of WesVarPC may also write to: Maida Montes, Westat, Inc., 1650 Research Boulevard, Rockville, MD 20850.

The Taylor series approach can also be used for the NHES. The two most commonly used software packages for this approach are SUDAAN and PCCARP. Both of these programs are for the PC and can be used to compute estimates of totals, means, and proportions as well as linear and logistic regression coefficients. Neither can account for nonresponse or raking adjustments to the weights, but for many estimates these adjustments are not critical for estimating the standard errors. The table in appendix C shows the proper design and nesting specifications to use in SUDAAN for each NHES data file.

SUDAAN is available through the Research Triangle Institute. Information on obtaining the software, including cost information, can be obtained by writing to Dr. Babu Shah, Research Triangle Institute, PO Box 12194, Research Triangle Park, NC 27709. Information on PCCARP, including costs, can be obtained by writing to Dr. Wayne Fuller, Department of Statistics, Iowa State University, Ames, IA 50010.

---

[2] WesVarPC will also import dBase and ASCII files.

An approach that is frequently used for complex analysis such as regression is to use the standard statistical software for exploratory data analysis and model fitting. Once the model is formulated, the appropriate analysis using WesVarPC or SUDAAN is used to estimate the parameters and the standard errors. This method is frequently used by analysts who are very familiar with a particular software package and feel more comfortable with using it during the exploratory stage. This is often a very reasonable compromise, since the final estimates are produced using the appropriate software. One disadvantage is that the exploratory analysis is done without the benefit of the fully valid estimates of the standard errors, but this is often not a problem in this type of analysis.

8.2.3 An Alternative Method for Producing Estimates and Standard Errors

Not all analysts will follow the recommended method of estimating standard errors. A common alternative approach and its likely consequences are discussed briefly below. In specific applications, this alternative may be valid and useful; however, for general purposes it has some shortcomings and may lead to statements that are not supported by the data.

This alternative method of analysis is to use a standard statistical package and then adjust the resulting standard errors by an average design effect. Most statistical software packages compute standard errors of the estimates based upon simple random sampling assumptions. The standard error from this type of statistical software can be adjusted for the complexity of the sample design to approximate the standard error of the estimate under the actual sample design used in the survey. For example, the variance of an estimated proportion in a simple random sample is the estimated proportion (p) times its complement (1-p) divided by the sample size (n). The standard error is the square root of this quantity. This estimate can be adjusted to more closely approximate the standard error for the estimates from the NHES.

A simple approximation of the impact of the sample design on the estimates of the standard errors of the estimates that has proved useful in NHES surveys and in many other surveys is to adjust the simple random sample standard error estimate by the root design effect (DEFT). The DEFT is the ratio of the standard error of the estimate computed using the replication method discussed above to the standard error of the estimate under the assumptions of simple random sampling. An average DEFT is computed by estimating the DEFT for a number of estimates and then averaging. In complex sample designs, like those used in the NHES, the DEFT is typically greater than unity due to the clustering of the sample and the differential weights attached to the observations. A standard error for an estimate can be approximated by multiplying the simple random sample standard error estimate by the average DEFT. Each NHES Data File User's Manual describes the calculation of the average design effect. The table in appendix C shows the average design effect for each NHES component. The next few paragraphs provide some examples of approximating standard errors using the average design effect.

Suppose that the NHES:95 ECPP data is used to obtain a weighted estimate of 60 percent for some characteristic (for example, suppose that 60 percent of children participate in some type of child care arrangement). An approximate standard error can be developed in a few steps. First, obtain the simple random sampling error for the estimate using the weighted estimate in the numerator and the unweighted sample size in the denomimator: the standard error for this 60 percent statistic would be the square root of ((60 x 40)/14,064) = 0.41, where the weighted estimate is 60 percent (p), 40 is 100 minus the estimated percent (100-p), and the unweighted sample size is 14,064 (n). The approximate standard error of the estimate from the NHES:95 ECPP data is this quantity (the simple random sample standard error) multiplied by the DEFT of 1.2. In this example, the estimated standard error would be 0.49 percent (1.2 x 0.41).

The approximate standard error for a mean can be developed using a related procedure. First, the mean is estimated using the full sample weight in a standard statistical package like SAS or SPSS. Second, the simple random sample standard error is obtained through a similar, but unweighted, analysis. Third, the standard error from the unweighted analysis is multiplied by the DEFT to approximate the standard error of the estimate under the NHES design. For example, suppose that in the NHES:95 ECPP data, the estimated (weighted) mean number of hours per week in nonparental care is 20 and the simple random sampling standard error (unweighted) is 5 hours. Then, the approximate standard error for the estimate would be 6 hours (5 x 1.2).

Users who wish to adjust the standard errors for parameter estimates of regression models should follow a procedure similar to that discussed for means, above. Specifically, the parameter estimates of the model can be estimated using a weighted analysis in a standard statistical software package such as SAS or SPSS. A similar, but unweighted, analysis will provide the simple random sample standard errors for these parameter estimates. The standard errors can then be multiplied by the DEFT to arrive at the adjusted standard error for the NHES design. For example, if a given variable in the NHES:95 ECPP data has a weighted estimate of 2.334 and an unweighted standard error of 0.45, then the adjusted standard error would be 1.2 x 0.45 = 0.54.

This method of approximating standard errors increases the standard errors of the estimates. Some research (Kish and Frankel 1974) suggests that the standard errors for more complex estimates such as regression coefficients may not be subject to design effects that are as large as that of statistics such as means and proportions. Thus, this alternative may lead to overestimating the standard errors of the estimates, but this error may be less problematic because it leads to confidence intervals and tests that still satisfy the nominal level.

References

Brick, J.M., and Waksberg, J. (1991). "Avoiding Sequential Sampling With Random Digit Dialing," *Survey Methodology* 17(1): 27-42.

Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). "Bias in List-Assisted Telephone Samples," *Public Opinion Quarterly* 59(2): 218-235.

Casady, R.J., and Lepkowski, J.M. (1993). "Stratified Telephone Survey Designs." *Survey Methodology* 19(1): 103-113.

Fortune, J. and McBee, J. (1984). "Considerations and Methodology for the Preparation of Data Files" in D.J. Bowering (Ed.) *Secondary Analysis of Available Data Bases, New Directions for Program Evaluation*. San Francisco: Jossy-Bass.

Kish, L. (1992). Weighting for Unequal Pi, *Journal of Official Statistics* 8: 183-200.

Kish, L. and Frankel, M. (1974). Inference from Complex Samples, *Journal of the Royal Statistical Society (B)* 36: 1-37.

Mohadjer, L., and West, J. (1992). *Effectiveness of Oversampling Blacks and Hispanics in the NHES Field Test*. National Household Education Survey Technical Report No. 5. (NCES Publication No. 92-104). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Rao, J.N.K., and Shao, J. (1992). "Jacknife Variance Estimation with Survey Data under Hot Deck Imputation." *Biometrika* 79: 811-822.

Rubin, D.R. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, NY.

Waksberg, J. (1978). "Sampling Methods for Random Digit Dialing," *Journal of the American Statistical Association* 73(361): 40-46.

APPENDIX A

NHES:91/93/95
Commonly Asked Questions and Answers

NHES:91/93/95
Commonly Asked Questions and Answers

Are the subjects in one NHES survey component related to subjects in other components?

The NHES is a repeated cross-sectional survey system, and each survey year uses an independent sample Therefore, the sample members from one year (e.g., the NHES:91) are independent from those another year (e.g., the NHES:93 or the NHES:95).

Some respondents or subjects within one survey year are related to one another. This is because households may have had members sampled for more than one component in a given survey cycle, or more than one household member may have been eligible for a single survey component.

For example:

- In the NHES:91, all eligible children were sampled. As a result, some subjects in the ECE component are siblings, cousins, or other relatives living within the same household.
- In the NHES:91, it is possible that household members were sampled for more than one component, that is, an adult may have been sampled for adult education and a child (or children) may have been sampled for ECE. The sampling rate for the AE component was reduced in households with sampled children in order to limit the response burden on the households.
- In the NHES:93, up to two children in the household were sampled for the SR component and up to two were sampled for the SS&D component. No child was sampled as the subject for more than one component.
- In the NHES:95, up to two children in each household were sampled for the ECPP component and up to two adults were sampled for the AE component. Thus, within a single household, up to four interviews may have been completed. However, sampling two adults within a household for AE interviews was relatively rare and only done in households containing adult education participants with less than a high school education.

The first 8 digits of the case identification number constitute the household identification, and can be used to identify multiple interviews within the same household.

What measures have been taken in NHES to insure against bias in the data?

About 94 percent of persons in the United States live in households with telephones. Those without telephones are different from those with telephones in some ways, notably in terms of their socioeconomic status. In the NHES, special weighting procedures are used to adjust the survey estimates to match totals from the Current Population Survey, using poststratification variables that are associated with telephone coverage. Additional information on telephone coverage bias is included in the Data File User's Manuals and in the NCES technical report *Telephone Undercoverage Bias of 14- to 21-year-olds and 3- to 5-year-olds* (NCES Publication No. 92-101). A forthcoming technical report will present research on the interruption of telephone service, that is, the extent to which households move in and out of telephone

coverage. This report is entitled, *Adjusting for Coverage Bias Using Telephone Service Interruption Data in the 1993 National Household Education Survey*.

Nonresponse is another important potential source of bias in any survey. The NHES project uses a calling protocol, refusal conversion efforts, and implementation of a Spanish language questionnaire to minimize bias resulting from unit (questionnaire) nonresponse. Item response for the NHES instruments is very high, more than 98 percent for nearly all items. Missing values were imputed for all items on the NHES:93 and NHES:95 public files; selected items were imputed in the NHES:91 files.

What do the negative numbers in the data sets mean?

The NHES data sets contain negative codes to designate missing data. In this way, it is possible to utilize the same missing data codes for all items regardless of the length of the field (number of columns) occupied by the data element. This enhances the uniformity of the file characteristics and simplifies the identification of missing values for data file users.

For all NHES data sets, a -1 (negative one) designates a legitimate skip, or cases for whom the variable is not appropriate. For example, if a person says that his/her child attends a public school, a -1 will appear in the item that asks if the school is affiliated with a religion, since the question is inappropriate.

The NHES:93 and NHES:95 data sets were fully imputed, so they do not contain other negative codes. In the NHES:91, however, only variables required for weighting or contributing to key derived variables were imputed. Other variables contain additional missing value codes. These are: -7, refused; -8, don't know, and -9, not ascertained. The last of these (-9) was assigned by data preparation staff during data cleaning and problem resolution; -7 and -8 were assigned during the interview.

What are the derived variables and how were they created?

Derived (or composite) variables are analytically useful constructs that are created using two or more variables. For example, an analyst may want to use the highest education of either parent in a household as a measure of socioeconomic status. The derived variable PARGRADE was constructed for this purpose, and appears in the Preprimary (NHES:91), Primary (NHES:91), School Readiness (NHES:93), School Safety and Discipline (NHES:93), and Early Childhood Program Participation (NHES:95) files. PARGRADE is composed of four variables: MOMGRADE (mother's highest grade), MOMDIPL (whether the mother has a high school diploma or equivalent), DADGRADE (father's highest grade), and DADDIPL (whether the father has a high school diploma or equivalent).

Some other derived variables are "counters," e.g., household size. The derived variables are designed to facilitate analysis by providing measures that are likely to be useful. The Data File User's Manuals include a discussion of each derived variable and its composition (chapter 6 of each User's Manual).

What if I'm interested in only a part of the population?

Many times, the research questions guiding one's analyses pertain to a specific population. That population may be defined by age, grade in school, family type, or other factors. Each of the NHES data files contains variables that can be used to subset the data file to include the population of interest.

A key variable provided for this purpose is called **MAINRSLT** (main interview result). This is a CATI system variable that defines the completion status of an interview, often defining a subpopulation to which the subject belongs. MAINRSLT can be used to subset the NHES:91 and NHES:95 Adult Education files, separating participants and nonparticipants. For the NHES:91 ECE component, MAINRSLT was used to create two separate data files -- Preprimary (preschoolers and kindergartners) and Primary (primary school students). In the School Readiness data set, MAINRSLT can be used to select preschoolers, kindergartners, primary students, and home schoolers. In School Safety and Discipline, MAINRSLT can be used to divide the data file between parents and youth, and by grade level (3rd through 5th, or 6th through 12th). In the NHES:95 Early Childhood Program Participation data set, MAINRSLT differentiates infants and toddlers, preschoolers, kindergartners, primary schoolers, and homeschoolers.

Another variable that can be used to subset some of the files is **ALLGRADE** (the enrollment status and grade of child). Analysts can use ALLGRADE to select only kindergartners from the Preprimary file, only 9th through 12th graders from the SS&D file, and so on. Each analyst will want to review the available measures in order to define the population of interest to him or her.

The **Electronic CodeBook (ECB)** program for the NHES facilitates efforts to subset the population. When preparing to write the extract program, the user is presented with a dialog box that allows him or her to define the extract population by age, race/ethnicity, sex, and enrollment status. By default, the extract population is *all* respondents in the catalog.

How do I calculate the standard errors for the NHES data?

The NHES sample designs are complex, multi-stage designs. As a result, it is erroneous to calculate standard errors for the estimates under simple random sampling assumptions. Each NHES data set includes two sets of variables that can be used to estimate the standard errors of statistics.

Replicate weights provided on each data file can be used to calculated standard errors using WesVarPC. This program uses replication methods to estimate standard errors.

PSU and *STRATUM* also appear on each data file and can be used in Taylor series approximations. There are software packages available for calculating standard errors using the Taylor series approach. Among these are SUDAAN and PCCARP.

See section 8 of this guide for additional discussion of variance estimation for the NHES data.

APPENDIX B

NHES:91/93/95 Information and Examples

Example 1: Checking a Subsetted File

This example shows ways in which a data user can check that his/her subsetting of a data file is correct (see section 5.2 for a discussion of subsetting NHES data files). For example, let's say that an analyst has subset the School Readiness file by selecting on MAINRSLT (the completion code for an interview), and selected preschoolers as the analysis population. There are two checks the analyst can do to be sure that the proper cases are included. Both checks involve comparing numbers from the file output and numbers from the codebook. In this example, we compare the total number of cases in the subfile with the total number of completed CN interviews in the codebook. As shown below, the number of cases in the subfile is 4,423 which matches the number of cases in the total file with a MAINRSLT value of CN.

Often, it is useful to select another variable to do a confirmatory check. For this example, we have selected ALLGRADE, a derived variable representing the child's enrollment status and grade. By definition, the derived variable ALLGRADE should equal 'not enrolled' or 'nursery/pre-k/Head Start' for all preschoolers (CNs). Comparing the SAS run from the extracted file with the codebook values for ALLGRADE, we see that the numbers of children enrolled in preschool (grade N) is the same (n = 2,084) and no children in the extract file have inappropriate values for ALLGRADE (e.g., kindergarten or a primary grade).

Check #1 -- Compare total number of CNs in MAINRSLT from the codebook with the total number of CNs in the SAS run from the subfile.

****

FROM ECB CODEBOOK

****

MAINRSLT = INTERVIEW COMPLETION STATUS

RECORD: 1 POSITION: 11-12
FORMAT: A2

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
CH COMPLETE HOME SCHOOL INTERVIEW	CH	62	0.6%	0.5%
CK COMPLETE KINDERGARTEN INTERVIEW	CK	2126	19.5%	19.7%
CN COMPLETE PRESCHOOL INTERVIEW	CN	4423	40.6%	42.8%
CS COMPLETE PRIMARY SCHOOL INTERVIEW	CS	4277	39.3%	37.0%
TOTALS:		10888	100.0%	100.0%

****

SAS OUTPUT

****

TABLE OF MAINRSLT BY ALLGRADE

MAINRSLT (INTERVIEW COMPLETION STATUS)	ALLGRADE (D-CHILD'S ENROLLMENT AND GRADE EQUIV)		
Frequency Percent Row Pct Col Pct	N NURS/PREK/HDST	0 NOT ENROLLED	TOTAL
CN COMPLETE PRE	2084 47.12 47.12 100.00	2339 52.88 52.88 100.00	4423 100.00
Total	2084 47.12	2339 52.88	4423 100.00

Check #2 -- Compare the number of preschoolers (CNs) in nursery/pre-k/Head Start programs in the codebook using ALLGRADE (n=2084) with the number of preschoolers in these programs in the SAS run from the subfile (n=2084). The number of preschoolers (CNs) who are not enrolled in the subfile (n=2339) does not match the total number of children not enrolled in ALLGRADE in the codebook (n=2340); however, this is not problematic. This is because all preschoolers in the subfile have appropriate ALLGRADE values and because children with ALLGRADE = 0 may have a MAINRSLT value other than CN. There is one case with ALLGRADE=0 and MAINRSLT=CK. This can be seen in a crosstabulation between ALLGRADE and MAINRSLT for all children (shown on next page). This child was reported to not be enrolled in school (ENROLL=2) and to not be in home school (HOMESCHL=2) and was assigned a MAINRSLT=CK instead of CN based on his/her age (AGE92=7).

****

FROM ECB CODEBOOK

****

ALLGRADE = D-CHILD'S ENROLLMENT AND GRADE/EQUIV

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
0 NOT ENROLLED	0	2340	21.5%	23.2%
N NURS/PREK/HDST	N	2084	19.1%	19.6%
T TRANS KIND	T	85	0.8%	0.8%
K KINDERGARTEN	K	2062	18.9%	19.1%
P PRE/TRANS FRST	P	13	0.1%	0.1%
1 1ST GRD/EQUIV	1	2147	19.7%	19.8%
2 2ND GRD/EQUIV	2	2140	19.7%	17.1%
3 3RD GRD/EQUIV	3	17	0.2%	0.3%
TOTALS:		10888	100.0%	100.0%

SAS OUTPUT

TABLE OF MAINRSLT BY ALLGRADE (all children)

MAINRSLT (INTERVIEW COMPLETION STATUS)	ALLGRADE (D-CHILD'S ENROLLMENT AND GRADE EQUIV)								
Frequency Percent Row Pct Col Pct	0 NOT ENROLLED	1 1ST GRD/ EQUIV	2 2ND GRD/ EQUIV	3 3RD GRD/ EQUIV	K KIND ERGAR	N NURS/ PREK/ HDST	P PRE/ TRANS FRST	T TRANS KIND	TOTAL
CH COMPLETE HOM	0 0.00 0.00 0.00	17 0.16 27.42 0.79	10 0.09 16.13 0.47	0 0.00 0.00 0.00	32 0.29 51.61 1.55	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 0.03 4.84 3.53	62 0.57
CK COMPLETE KIN	1 0.01 0.05 0.04	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2030 18.64 95.48 98.45	0 0.00 0.00 0.00	13 0.12 0.61 100.00	82 0.75 3.86 96.47	2126 19.53
CN COMPLETE PRE	2339 21.48 52.88 99.96	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2084 19.14 47.12 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4423 40.62
CS COMPLETE PRI	0 0.00 0.00 0.00	2130 19.56 49.80 99.21	2130 19.56 49.80 99.53	17 0.16 0.40 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4277 39.28
TOTAL	2340 21.49	2147 19.72	2140 19.65	17 0.16	2062 18.94	2084 19.14	13 0.12	85 0.78	10888 100.00

Example 2: Missing Values

As discussed in section 7, full imputation was performed on the NHES:93 and NHES:95 data sets. Therefore, all missing values (denoted by the code -1) are legitimate skips. A legitimate skip can occur for two reasons. First, answers to one or more previously asked question(s) can result in a skip. For example, if a child attends a public school, the question about whether a private school was affiliated with a religion would equal -1. This example is illustrated by the SAS run below on NHES:93 School Readiness data.

****

SAS RUN

****

Table of PPUBL by MAINRSLT

PPUBL (R71 - Current school public or private)
MAINRSLT (Interview completion status)

Frequency Percent Row Pct Col Pct	CH	CK	CN	CS	Total
-1 Inapplicable	62 0.57 0.94 100.00	2126 19.53 32.16 100.00	4423 40.62 66.90 100.00	0 0.00 0.00 0.00	6611 60.72
1 Public	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3750 34.44 100.00 87.68	3750 34.44
2 Private	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	527 4.84 100.00 12.32	527 4.84
Total	62 0.57	2126 19.53	4423 40.62	4277 39.28	10888 100.00

Of the primary school path (CS) respondents who were administered this item (PPUBL) in the School Readiness interview, 527 answered that the child attended a private school and were subsequently asked the next item (PCHURCH). All other respondents (n = 10,361) were assigned a value of -1 for PCHURCH either because their child attends a public school or because their child is not in primary school (CH, CK, and CN).

Table of PCHURCH by MAINRSLT

PCHURCH (R73-Religion-affiliated school)
MAINRSLT (Interview completion status)

Frequency Percent Row Pct Col Pct	CH	CK	CN	CS	Total
-1 Inapplicable	62 0.57 0.60 100.00	2126 19.53 20.52 100.00	4423 40.62 42.69 100.00	3750 34.44 36.19 87.68	10361 95.16
1 Religion-affiliated	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	427 3.92 100.00 9.98	427 3.92
2 Not religion-affiliated	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	100 0.92 100.00 2.34	100 0.92
Total	62 0.57	2126 19.53	4423 40.62	4277 39.28	10888 100.00

An entire block of items can be set equal to -1 if the items are not applicable to the interview path. As illustrated by the SAS run below, only preschoolers (MAINRSLT = CN) are asked items from the School Readiness developmental profile series; the other paths (CH, CK, and CS) are set to -1 for all items in the section (n = 6465).

Table of DPPENCIL by MAINRSLT

DPPENCIL (R19-CHILD HOLDS PENCIL PROPERLY)
MAINRSLT (INTERVIEW COMPLETION STATUS)

Frequency Percent Row Pct Col Pct	CH	CK	CN	CS	Total
-1 Inapplicable	62 0.57 0.96 100.00	2126 19.53 32.88 100.00	0 0.00 0.00 0.00	4277 39.28 66.16	6465 59.38
1 Yes	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4013 36.86 100.00 90.73	0 0.00 0.00 0.00	4013 36.86
2 No	0 0.00 0.00 0.00	0 0.00 0.00 0.00	410 3.77 100.00 9.27	0 0.00 0.00 0.00	410 3.77
Total	62 0.57	2126 19.53	4423 40.62	4277 39.28	10888 100.00

Example 3: Comparison of wordings and response categories for similar items in NHES:91, NHES:93, and NHES:95

Section 4.2 discussed some potential pitfalls in data preparation, including "variant variables." This example presents frequencies for some items appearing in the NHES:91, NHES:93 and NHES:95. Note that item wording and categories are not identical. When comparing similar items from multiple survey administrations, the researcher should carefully compare the wording of each item and the population of whom the item was asked.

In the NHES:91, two versions of a reading question appeared. The first asked about *general* reading frequency, and was not tied to a specific time period. The second asked about reading to the child in the week prior to the interview.

****

NHES:91 VERSION 1

****

READTO -- P19 HOW OFTEN READ TO CHILD

P19. About how often do you (and OTHER PARENT/GUARDIAN) read stories to (CHILD)?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
NEVER	1	74	1.0%	1.0%
SEVERAL TIMES A YEAR	2	180	2.4%	2.4%
SEVERAL TIMES A MONTH	3	1076	14.1%	14.8%
AT LEAST THREE TIMES A WEEK	4	2926	38.2%	38.8%
EVERY DAY	5	3390	44.3%	43.0%
RESERVED CODES:				
DK	-8	7	0.1%	(MISS)
NOT ASCERTAINED	-9	2	0.0%	(MISS)
TOTALS:		7655	100.00%	100.00%

NOTE: Item P19 is from the NHES:91 Preprimary data file. This variable also appears in the NHES:91 Primary data file as question E36.

READTO -- P24 READ TO CHILD LAST WK

P24. In the past week, have you or someone in your family done the following this with (CHILD)? Read to (him/her)?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
YES	1	7218	94.3%	94.2%
NO	2	414	5.4%	5.8%
RESERVED CODES:				
DON'T KNOW	-8	11	0.1%	(MISS)
NOT ASCERTAINED	-9	12	0.2%	(MISS)
TOTALS:		7655	100.0%	100.0%

WKREADN -- P24 # TIMES READ TO CHILD IN PAST WK

P24. How many times? Would you say one or two times or three or more?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
ONE OR TWO TIMES	1	1676	21.9%	24.0%
THREE OR MORE TIMES	2	5539	72.4%	77.0%
RESERVED CODES:				
INAPPLICABLE	-1	437	5.7%	(MISS)
DON'T KNOW	-8	3	0.0%	(MISS)
TOTALS		10888	100.0%	100.0%

NOTE: Item P24 is from the Preprimary file. This item also appears on the Primary file as question E44.

In the NHES:93 SR component, two versions of a reading question were asked again. Both items focussed on the numbers of times the child had been read to in the week prior to the interview. One was a single item, and the other was a three-stage item. Each was asked of a split-half sample of respondents.

****

NHES:93 -- VERSION 1

****

READTIME = R96A-TIME FAMILY READ TO CHILD LAST WK

R96A. Now I'd like to talk with you about activities in your home in the past week. How many times have you or someone in your family read to (CHILD) in the past week? Would you say. . . not at all, once or twice, three or more times, or every day?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
1 NOT AT ALL	1	431	4.0%	7.7%
2 ONCE OR TWICE	2	1048	9.6%	19.6%
3 3 OR MORE TIMES	3	15559	14.3%	29.0%
4 EVERY DAY	4	2359	21.7%	43.7%
RESERVED CODES				
-1 INAPPLICABLE	-1	5491	50.4%	(MISS)
TOTALS:		10888	100.0%	100.0%

****

NHES:93 -- VERSION 2

****

READTO -- R96-FAMILY MEMBER READ TO CHILD LAST WK

R96. Now I'd like to talk with you about activities in your home in the past week. In the past week, have you or has someone in your family read to (CHILD)?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
1 YES	1	4926	45.2%	90.3%
2 NO	2	565	5.2%	9.7%
RESERVED CODES:				
-1 INAPPLICABLE	-1	5397	49.6%	(MISS)
TOTALS:		10888	100.0%	100.0%

READTON -- R97-TIMES/WK FAMILY READ TO CHILD

R97. How many times? Would you say . . .one or two times or three or more times?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
1 ONE OR TWO TIMES	1	1118	10.3%	22.4%
2 THREE OR MORE TIMES	2	3808	35.0	77.6%
RESERVED CODES:				
-1 INAPPLICABLE	-1	5962	54.8%	(MISS)
TOTALS		10888	100.0%	100.0%

READDAY -- R98-READING EVERY DAY IN LAST WEEK

R98. Was that every day in the past week?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
1 YES	1	2563	23.5%	6.6%
2 NO	2	1245	11.4%	32.4%
RESERVED CODES:				
-1 INAPPLICABLE	-1	7080	65.0%	(MISS)
TOTALS:		10888	100.0%	100.0%

In the NHES:95 ECPP component, there was a single item asking about the frequency with which family members read to the child in the past week.

****

NHES:95

****

HAREADFM = L1-TIMES FAMILY READ TO CHILD LAST WK

L1. How many times have you or someone in your family read to (CHILD) in the past week? Would you say. . . not at all, once or twice, three or more times, or every day?

RESPONSE	CODES	FREQ	UNWGTD PERCENT	WGTD PERCENT
1 NOT AT ALL	1	1360	9.7%	9.7%
2 ONCE OR TWICE	2	2458	17.5%	17.6%
3 3 OR MORE TIMES	3	3496	24.9%	24.4%
4 EVERY DAY	4	6750	48.0%	48.3%
TOTALS:		14064	100.0%	100.0%

APPENDIX C

NHES:91/93/95 Summary of Weighting and Sample Variance Estimation Variables

Summary of Weighting and Sample Variance Estimation Variables

NHES Data File	Full Sample Weight	Computing Sampling Errors				Approximating Sampling Errors
		Replication Method (WesVarPC)		Taylor Series Method (SUDAAN)		DEFT (Average Root Design Effect)
		Replicate Weights	Jackknife Method	Sample Design	Nesting Variables	
NHES:91 *Early Childhood Education*						
■ Primary file	EWGT	EWREPL1 - EWREPL50	JK1	WR	VSTRAT PSU	1.3
■ Preprimary file	EWGT	EWREPL1 - EWREPL50	JK1	WR	VSTRAT PSU	1.3
NHES:91 *Adult Education*						4.5 Full Sample
■ Adult file	AEWT	AEREPL1 - AEREPL50	JK1	WR	VSTRAT PSU	2.3 Participants
■ Course file[1]	AEWT	AEREPL1 - AEREPL50	JK1	WR	VSTRAT PSU	2.8 Nonparticipants 3.8 Blacks 3.2 Hispanics 2.8 White (non-Hispanic) 2.4 Other races
NHES:93 *School Readiness*	FWGT0	FWGT1 - FWGT60	JK2	WR	STRATUM PSU	1.3
NHES:93 *School Safety & Discipline*						
■ Parent interviews only	FWGT0	FWGT1-FWGT60	JK2	WR	STRATUM PSU	1.4
■ Parent & Emancipated Youth (EY) interviews	FWGT0 (for parents) & PFWGT0 (for EY)	FWGT1-FWGT60, PFWGT1-PFWGT60	JK2	WR	STRATUM PSU	1.4
■ Youth interviews (including Emancipated Youth)	FWGT0	FWGT1-FWGT60	JK2	WR	STRATUM PSU	1.5
NHES:95 *Early Childhood Program Participation*	EWEIGHT	ERPL1 - ERPL50	JK1	WR	STRATUM PSU	1.2
NHES:95 *Adult Education*[2]	AEWEIGHT	ARPL1 - ARPL50	JK1	WR	STRATUM PSU	1.3

[1] Unlike the NHES:95 Adult Education data file, no course weights are provided in the NHES:91 course file. The full sample weight and variables for computing sampling errors are provided in the course file for making adult-level estimates. Information as to the total number of courses that adults took is also available, and procedures similiar to those described in the NHES:95 Adult Education Data File User's Manual could be used to create weights for making course-related estimates. However, it is important to note that the course information collected in the NHES:91 pertains to the four most recent courses taken, rather than a random sample of courses as was the case in the NHES:95.

[2] This data file contains weights for making "person-course" estimates pertaining to work-related and other formal structured courses. Weights are required for these types of courses because course-related data were collected only for a random subsample of courses. The weight variables are called WRWGT and SAWGT. A simple way of doing this is to create a new variable that is the product of the course weight and the variable of interest. The standard weight and variance estimation methods are then applied to the new variable. See the NHES:95 Adult Education Data File User's Manual for more details.

BEST COPY AVAILABLE

United States
Department of Education
Washington, DC 20208–5651

Official Business
Penalty for Private Use, $300

Postage and Fees Paid
U.S. Department of Education
Permit No. G–17

Standard Mail (B)